SRI International

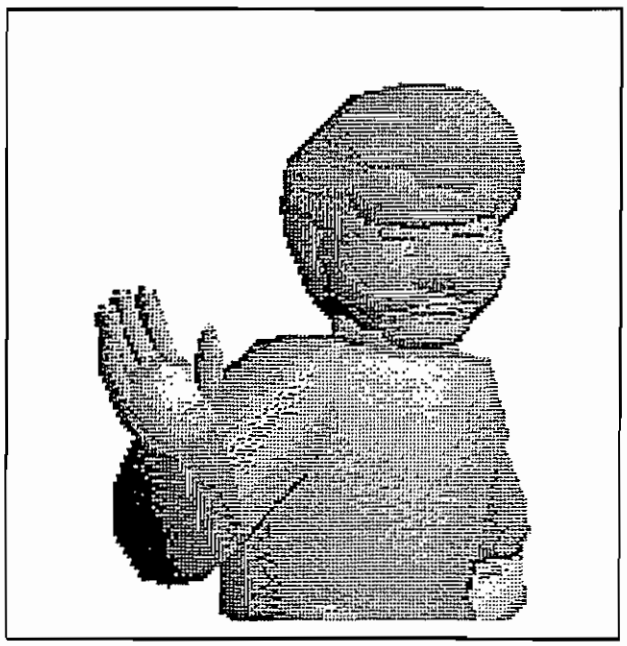
# PERCEPTUAL ORGANIZATION AND THE REPRESENTATION OF NATURAL FORM

**Technical Note No. 357 (Revised)**

July 29, 1986

**By:** Alex P. Pentland, Computer Scientist

Artificial Intelligence Center
Computer Science and Technology Division



SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **29 JUL 1986** | | **00-07-1986 to 00-07-1986** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Perceptual Organization and the Representation of Natural Form** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | | **42** | |

# Perceptual Organization and the Representation of Natural Form

**Alex P. Pentland**
*Artificial Intelligence Center, SRI International,*
*333 Ravenswood Ave, Menlo Park, CA 94025, U.S.A.*
*Center for the Study of Language and Information,*
*Stanford University, Stanford, CA 94038, U.S.A.*

Recommended by Daniel G. Bobrow and Pat Hayes

ABSTRACT

*To support our reasoning abilities perception must recover environmental regularities—e.g., rigidity, "objectness," axes of symmetry—for later use by cognition. To create a theory of how our perceptual apparatus can produce meaningful cognitive primitives from an array of image intensities we require a representation whose elements may be lawfully related to important physical regularities, and that correctly describes the perceptual organization people impose on the stimulus. Unfortunately, the representations that are currently available were originally developed for other purposes (e.g., physics, engineering) and have so far proven unsuitable for the problems of perception or common-sense reasoning. In answer to this problem we present a representation that has proven competent to accurately describe an extensive variety of natural forms (e.g., people, mountains, clouds, trees), as well as man-made forms, in a succinct and natural manner. The approach taken in this representational system is to describe scene structure at a scale that is similar to our naive perceptual notion of "a part," by use of descriptions that reflect a possible formative history of the object, e.g., how the object might have been constructed from lumps of clay. For this representation to be useful it must be possible to recover such descriptions from image data; we show that the primitive elements of such descriptions may be recovered in an overconstrained and therefore reliable manner. We believe that this descriptive system makes an important contribution towards solving current problems in perceiving and reasoning about natural forms by allowing us to construct accurate descriptions that are extremely compact and that capture people's intuitive notions about the part structure of three-dimensional forms.*

## 1. Introduction

Our world is very highly structured: evolution repeats its solutions whenever possible [1], and inanimate forms are constrained by physical laws to a limited number of basic patterns [2]. The apparent complexity of our environment is produced from this limited vocabulary by compounding these basic forms in myriad different combinations. Indeed, the highly patterned nature of our

*Artificial Intelligence* **28** (1986) 293–331
0004-3702/86/$3.50 

environment is a necessary precondition for intelligence; for if the apparent complexity of our environment were approximately the same as its intrinsic (Kolmogorov) complexity, then intelligent prediction and planning would be impossible, for there would be no lawful relations. It is this internal structuring of our environment, then, that causes object features to cluster into groups, and allows us to reason successfully using the simplified category descriptions that we typically employ [3].

To support our reasoning abilities, therefore, perception must recover these environmental regularities—e.g., rigidity, "objectness," axes of symmetry—for later use in cognitive processes. This recovery of structure is known as *perceptual organization*, familiar from such research efforts as the Gestalt movement [4], Johansson's [5] study of the organization of motion perception, and more recently Marr and Nishihara's [6, 7] theory of form perception using a description based on generalized cylinders [8]. The problem of perceptual organization is important because the structural regularities that perception recovers are the parts from which we construct our picture of the world; they are the building blocks of all cognitive activities.

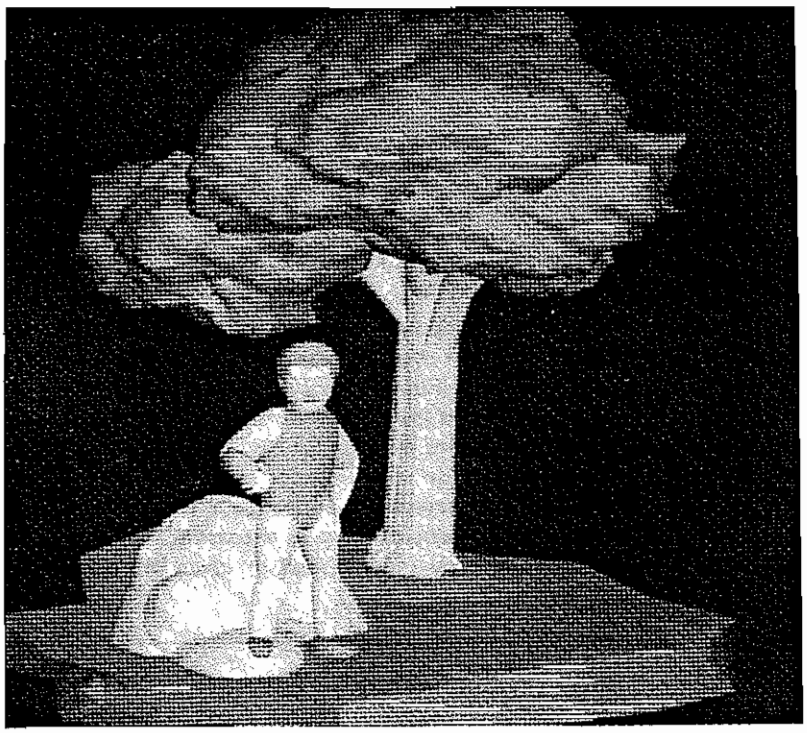


FIG. 1. A scene described and generated by the representational system described within: tree leaves and bark, rocks and hair are fractal surfaces, the overall shape is described by Boolean combination of appropriately deformed superquadrics. Only 56 primitives are required (fewer than 500 bytes of information) to specify this scene. The slightly cartoon-like appearance is primarily due to the lack of surface texturing.

In this paper we will approach the problem of visual perceptual organization in a manner similar to Gibson [9] or Marr [10]: we want to construct a theory of how our perceptual concepts—shape, objectness, and the like—are lawfully related to the regularities (structure) in our environment. Like Marr, but unlike Gibson, we desire a *computational* theory: one that details how the physical regularities of our world allow the computation of meaningful descriptions of our surroundings by use of image data. Further, we want these descriptions to match the perceptual organization we impose on the stimulus—the one structuring of the stimulus that we know can support general-purpose cognitive activity. That is, we want a theory that both details how meaningful assertions can be derived from image data and accounts for human perceptual characteristics.

The core of any such theory must be a representation that is isomorphic to our perceptual organization, and whose elements can be computed from the unstructured array of image intensities. Unfortunately, the representations that are currently available were originally developed for other purposes (e.g., the pointwise descriptions of physics or the platonic-solids descriptions of engineering) and are therefore often unsuitable for the problems of perception.

Most current-day vision research, for instance, is based on the pointwise representation used in describing the physics of image formation, and consequently research has focused on analyzing image content on a local, point-by-point basis. Biological visual systems, however, can not recover scene structure from such local information.[1] In fact, biological visual systems are strikingly insensitive to the point-by-point particulars of the image-formation process (e.g., reflectance function or illuminant direction), factors that figure prominently in today's best vision research.

Rather than depending only upon pointwise information, people seem to make heavy use of the larger-scale structure of the scene in order to guide their perceptual interpretation. Similarly, the performance of most current-day vision algorithms depends critically upon assumed larger-scale structural context, e.g., upon assuming smoothness or isotropy. To progress towards general-purpose vision, therefore, we need new representations capable of describing these critical larger-scale structures; the "parts" or "building blocks" that we use to organize the image and provide a framework for perceptual interpretation.

Towards this end Marr and Nishihara [6] proposed a scheme using hierarchies of cylinder-like modeling primitives to describe natural forms. Their proposal is, it seems, the most widely known representation suggested to date; it captures many of our intuitions about axes of symmetry and hierarchical description (see also [11–14]). Further, in recent years representations like

[1] As you can confirm for yourself by looking through a long, one-inch wide tube such as found in rolls of wrapping paper.

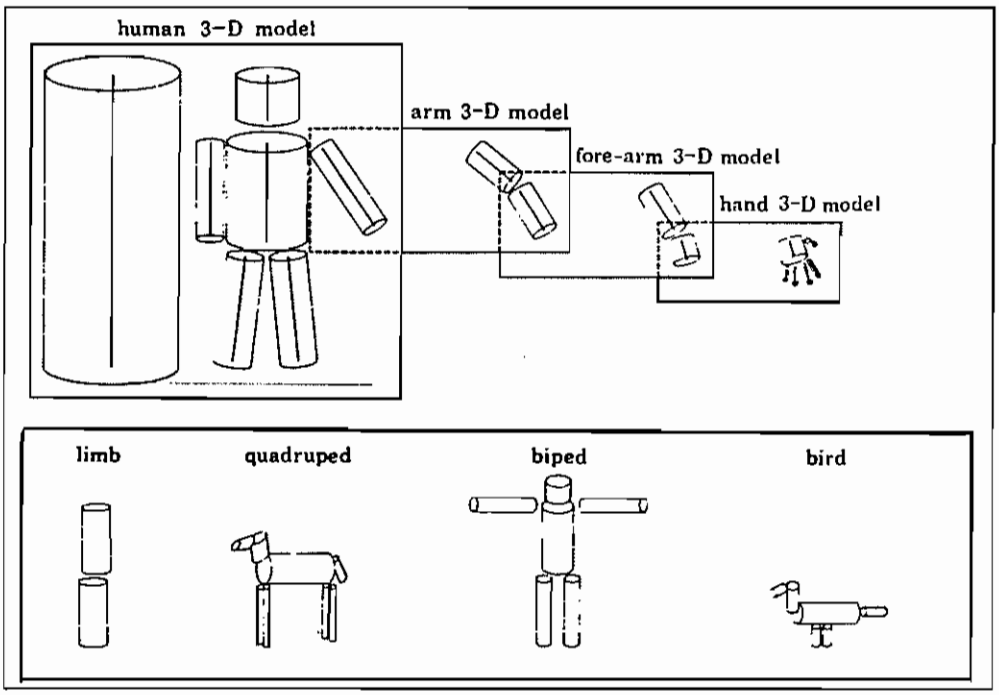



FIG. 2. Marr and Nishihara's scheme for the description of biological form.

theirs have found considerable success in industrial-style machine-vision systems where an exact model of the specific objects that are to be discovered in the image data is available [15, 16]. Unfortunately, such a representation is only capable of an extremely abstracted description of most natural and biological forms, as is illustrated in Fig. 2. It cannot accurately and succinctly[2] describe most natural animate forms or produce a succinct description of complex inanimate forms such as clouds or mountains.

In this paper we will present a representational system that has proven competent to accurately describe an extensive variety of natural forms (e.g., people, mountains, clouds, trees), as well as man-made forms, in a succinct and natural manner. Fig. 1 shows an example of a scene described in this representation; only 56 descriptive "parts" (about 500 bytes of information) were employed. We will then present evidence that we can use the special properties of this representational system to recover descriptions of specific objects from image data, and finally we will argue that these recovered descriptions are extremely useful in supporting both common-sense reasoning and man–machine communication.

[2] If we retreat from cylinders to generalized cylinders we can, of course, describe such shapes accurately. The cost of such retreat is that we must introduce a 1-D (at least) function describing the sweeping function; which makes the representation neither succinct nor intuitively attractive.

## 2. Vision, Cognition, and Models of Scene Structure

Perception is the mind's window on the world: its task is to recognize and report objects and relations that are important to the organism. It is this perceptual link between the *objective* environment and our *conception* of the environment that makes our thoughts meaningful; it ensures that they have some correspondence with the surrounding world.

Because the objects and relations recovered by perception are the primitives upon which all cognition is built, the particular way in which our perceptual apparatus organizes sensory data—that is, which regularities are noted and which are ignored—places strong constraints on the ways in which we can think about our environment. When perception organizes the sensory data in a way unsuited to the current task even simple problems can become nearly impossible to solve, as is illustrated by problems where you "see" the solution only when you "look" at them in the right way.

Identifying important environmental regularities and relating them to the primitive elements of cognition is, therefore, crucial to an understanding of cognitive function, and has consequently become the principal goal of research into visual function [9, 10]. The central problem in such research is, of course, that the sensory data underdetermines the scene structure. Image pixels, by themselves, can determine nothing. Some model of image formation and environmental structure is *required* in order to obtain any assertion about the viewed scene.

To construct a theory relating cognitive primitives to environmental structure, therefore, we must view visual perception as the process of recognizing image regularities that are known—on the basis of one's model of the world—to be reliably and lawfully related to cognitive primitives. The need for a model cannot be sidestepped, for it is the model that relates the theory's representations and computations to the state of the real world, and thus explains the semantics—the *meaning*—of the theory. A theory of visual function that has no model of the world also has no meaning.[3]

Understanding the early stages of perception as the interpretation of sensory data by use of models (knowledge) of the world has, of course, become a standard vision research paradigm. To date, however, most models have been of two kinds: high-level, *specific* models, e.g., of people or houses, and

[3] Theories of visual function, therefore, are based on models: models of how the world is structured and of how this structure is evidenced by regularities in the image. Much vision research is *not* model-based, of course: research on the mechanisms of vision (e.g., parallel processors, neurons), or on procedures for accomplishing visual tasks (e.g., variational calculus, relaxation methods) need not employ models of the world. But to understand visual *function*—that is, how one can infer information about the world—it is necessary to have a model of the salient world structure and of how that structure evidences itself in the image. Only then can one understand how certain features of the image can allow recognition and recovery of the information of interest.

low-level models of image formation, e.g., of edges. The reason research has almost exclusively focused on these two types of model is a result more of historical accident than concious decision. The well-developed fields of optics, material science and physics (especially photometry) have provided well worked-out and easily adaptable models of image formation, while engineering, especially recent work in computer-aided design, have provided standard ways of modeling industrial parts, airplanes and so forth.

Both the use of image-formation models and specialized models has been thoroughly investigated. It appears to us that both types of models, although useful for many applications, encounter insuperable difficulties when applied to the problems faced by, for instance, a general-purpose robot. In the next two subsections we will examine both types of models and outline their advantages and disadvantages for recovering important scene information. In the remainder of this section we will then motivate, develop and investigate an alternative category of models.

### 2.1. Model of image formation

Most recent research in computational vision has focused on using pointwise models of image formation borrowed from optics, material science and physics. This research has been pursued within the general framework originally suggested by Marr [10] and by Barrow and Tenenbaum [17], in which vision proceeds through a succession of levels of representation. The initial level is computed directly from local image features, and higher levels are then computed from the information contained in small regions of the preceding levels. Processing is primarily data-driven (i.e., bottom-up).

In Marr's scheme the initial level is called the "raw primal sketch," and contains a description of significant local image structure, e.g., edges, lines, or flowfield vectors, represented in the form of an array of feature descriptors that preserves the local two-dimensional geometry of the image. The second level is called the "$2\frac{1}{2}$-D sketch," and is intended to describe local surface properties (e.g., color, orientation) and discontinuities in a viewer-centered coordinate frame. Again, the recovered local surface properties are placed in a set of numeric arrays in registration with original image. From this point an object-centered, volumetric representation was to be computed, such as is illustrated by Fig. 2. The rationale for this level of representation is that tasks such as navigation or object recognition seem to require description in a viewpoint-independent coordinate frame.

Despite its prevalence, there are serious problems that seem to be inherent to this research paradigm. Because scene structure is underdetermined by the local image data [18], researchers have been forced to make *unverifiable* assumptions about large-scale structure (e.g., smoothness, isotropy) in order to derive useful information from their local analyses of the image. In the real

world, unfortunately, such assumptions are often seriously in error: in natural scenes the image-formation parameters change in fairly arbitrary ways from point to point, making any assumption about local context quite doubtful. As a result, those techniques that rely on strong assumptions such as isotropy or smoothness have proved fragile and error-prone; they are simply not useful for many natural scenes (for a more extended discussion see Witkin and Tenebaum [19]).

That such difficulties have been encountered should not, perhaps, be too surprising. It is easily demonstrated (by looking through a viewing or reduction tube) that people can obtain little information about the world from a local image patch taken out of its context. It is also clear that detailed, analytic models of the image-formation process are not essential to human perception; humans function quite well with range finder images (where brightness is proportional to distance rather than a function of surface orientation), electron-microscope images (which are approximately the reverse of normal images), and distorted and noisy images of all kinds—not to mention paintings and drawings.

Perhaps even more fundamentally, however, even if depth maps and other maps of intrinsic surface properties could be reliably and densely computed, how useful would they be? As Witkin and Tenenbaum point out, industrial vision work [16] using laser range data has demonstrated that the depth maps, reflectance maps and the other maps of the $2\frac{1}{2}$-D sketch are still basically just images. Although useful for obstacle avoidance and other very simple tasks, they still must be segmented, interpreted and so forth before they can be used for any more sophisticated task. The conclusion to be drawn from such work is that image-like measurements of range and other surface properties contribute incrementally, in much the same way as color: they add a dimension that simplifies some decisions, but they do *not* solve the difficult problems encountered in image interpretation.

### 2.2. Specialized models

The alternative to models of image formation has been engineering-style representations; e.g., CAD-CAM models of specific objects that are to be identified and located. Such detailed, specific models evidence themselves in image data in an extremely complex manner, in part because the models themselves are often complex, but more importantly because it is the object's surface shape, and not the appearance of the object, that is described. As the object's orientation varies, therefore, these models produce a *very* large number of different pixel configurations—to say nothing of what happens when we vary the illumination and imaging conditions. As a consequence, the image regularities that allow reliable recognition across all of the allowable configurations are very subtle and complex.

The large number of possible appearances for such models makes the problem of recognizing them very difficult—unless an extremely simplified representation is employed. The most common type of simplified representation is that of a wireframe model whose components correspond to the imaged edges. Such a simplied representation permits reliable recognition of models with currently available computational resources, given that we are in a restricted environment where the descriptive power of such wireframe models is sufficient, e.g., as in industrial applications. As a result systems based on CAD-like models of specific objects have provided most of the success stories in machine vision.

Despite this success, the use of an impoverished representation generally means that the flexibility, reliability and discriminability of the recognition process is limited. Thus research efforts employing specific object models have floundered whenever the number of objects to be recognized becomes large, when the objects may be largely obscured, or when there are many unknown objects also present in the scene.

An even more substantive limitation of systems that employ *only* high-level, specific models is that there is no way to learn new *types* of objects: new model types must be specially entered, usually by hand, into the database of known models. This is a significant limitation, because the ability to encounter a new type of object, enter it into a catalog of known objects, and thereafter recognize it is an absolute requirement of truly general-purpose vision.

### 2.3. Part and process models

Some sort of additional constraint is required to oversome the fundamental problem of insufficient information being available from the image. If sufficient constraint is not available from models of image formation, then from where? Human vision seems to function quite well as long as the imaging process preserves the basic spatial structure of the scene. It seems, therefore, that human perception must be exploiting constraints provided by the structure of the scene without reliance on quantitative, pointwise models of the image-formation process. What is required, then, are models of scene structure that capture something about the larger-scale structure of our environment. We cannot, however, appeal to CAD-like models of specific objects because of the impossibility of learning new descriptions.

In response to these seemingly intractable problems some researchers have begun to search for a third type of model, one with a grain size intermediate between the pointwise models of image formation and the complex, specific models of particular objects (see [20]). There is good reason to believe that it may be possible to accurately describe our world by means of such intermediate-grain models; that world can be modeled as a relatively small set of generic processes that occur again and again, with the apparent complexity of

our environment being produced from this limited vocabulary by compounding these basic forms in myriad different combinations.

We have known for over a century that evolution repeats its solutions whenever possible [1], resulting in great regularities across all species: there are but a few types of limb, a few types of skin, a few types of leaf, and a few patterns of branching. An amazingly good model of a tree, for instance, is the composition of a simple branching process with three-dimensional texture processes for generating bark and leaves [21]; the same branching models can also serve for rivers, veins, or coral. Similarly, it is now being discovered that inanimate forms may also be constrained by physical laws to a limited number of basic patterns [2,22]. Mandelbrot has shown that such apparently complex forms such as clouds, hills, coastlines or cheese can all be described by simple patterns recursively repeated at all different scales [22], while Stevens presents strong evidence that natural textures occur in but a few basic forms [2].

It is this internal structuring in our environment that allows us to derive lawful relationships [23].[4] It is exactly this internal structuring of our environment that causes object features to cluster into groups, and thus allows us to successfully employ simplified category descriptions for common-sense reasoning [3].

It appears, then, that it may be possible to accurately model the world in terms of *parts*: macroscopic models that, in relatively simple combination, can be used to form rough-and-ready models of the objects in our world and how they behave. If we adopt this view, then the central problem of perception research is *not* Marr's scheme of successively describing images, surfaces, and volumes, with the hope that we will eventually arrive at recognition of high-level models [10]. Rather, the central problem for perception is to find a set of generically applicable part models, discover image regularities that are lawfully associated with the individual parts, and then use these regularities to recognize the content of an image as a combination of these generic primitives. This new proposal, then, is that our theory of perception can dispense entirely with these initial stages of description and begin immediately with recognition of part models: models that are in principle much like models of houses and chairs, but that are more generally applicable and less detailed.

Because such models would be simpler than models of specific objects we would expect that we could more readily characterize how they would appear in an image. On the other hand, because they describe larger-scale structure than pointwise models of image formation, we would expect that they might not suffer from the problems of underdetermination that have forced researchers to make unrealistically strong assumptions such as smoothness or isotropy. Besides offering a good balance between complexity and reliability, such

[4] If the apparent complexity of our environment were equal to its intrinsic Kolmogorov complexity, then no lawful relationships would be possible.

intermediate-grain part models spark considerable interest because they describe the world in the right terms: they speak qualitatively of whole objects and of relations between objects, rather than of local surface patches or of specific objects. Thus, they can potentially provide a vocabulary for describing the world at the grain size that is most often directly useful to us.

The problem with forming such "part" models is that they must be complex enough to be reliably recognizable, and yet simple enough to reasonably serve as building blocks for specific object models. Current 3-D machine-vision systems, for instance, typically use rectangular solids and cylinders to model specific shapes. Using these primitives for the automatic construction of a description for an arbitrary new object has not proven possible, except[5] (as in industrial or urban imagery) when the set of objects that will be encountered is constrained to be simple combinations of rectangular solids or cylinders [24]. To support truly general-purpose vision, therefore, we need to develop new modeling primitives that can be used to build descriptions of arbitrary objects and that are recognizable in standard imagery. Our work towards this goal is the subject of the remainder of this paper.

## 3. A Representation for Natural Forms

We present here a representational system that has been proven competent to accurately describe an extensive variety of natural forms (e.g., people, mountains, clouds, trees), as well as man-made forms, in a succinct and natural manner (see Fig. 1). The idea behind this representational system is to provide a vocabulary of models and operations that will allow us to model our world as the relatively simple composition of component "parts," parts that are reliably recognizable from image data.

The most primitive notion in this representation may be thought of as a "lump of clay," a modeling primitive that may be deformed and shaped, but which is intended to correspond roughly to our naive perceptual notion of "a part." It is worth noting that this notion of "part" agrees with that used by Konderink and Van Doorn [25, 26] or by Hoffman and Richards [27] in their analysis of how part boundaries impose constraints upon three-dimensional surfaces, although they did not actually propose a model of what constitutes a three-dimensional "part." For this basic modeling element we use a parameterized family of shapes known as a *superquadrics* [28, 29], which are described (adopting the notation $\cos \eta = C_\eta$, $\sin \omega = S_\omega$) by the following equation:

[5] A caveat should be noted with respect to laser range finders and the like: in some cases the thousands of range measurements provided by these active sensors can give enough additional constraint to allow recovery of low-level, polygon-like descriptions of novel objects.

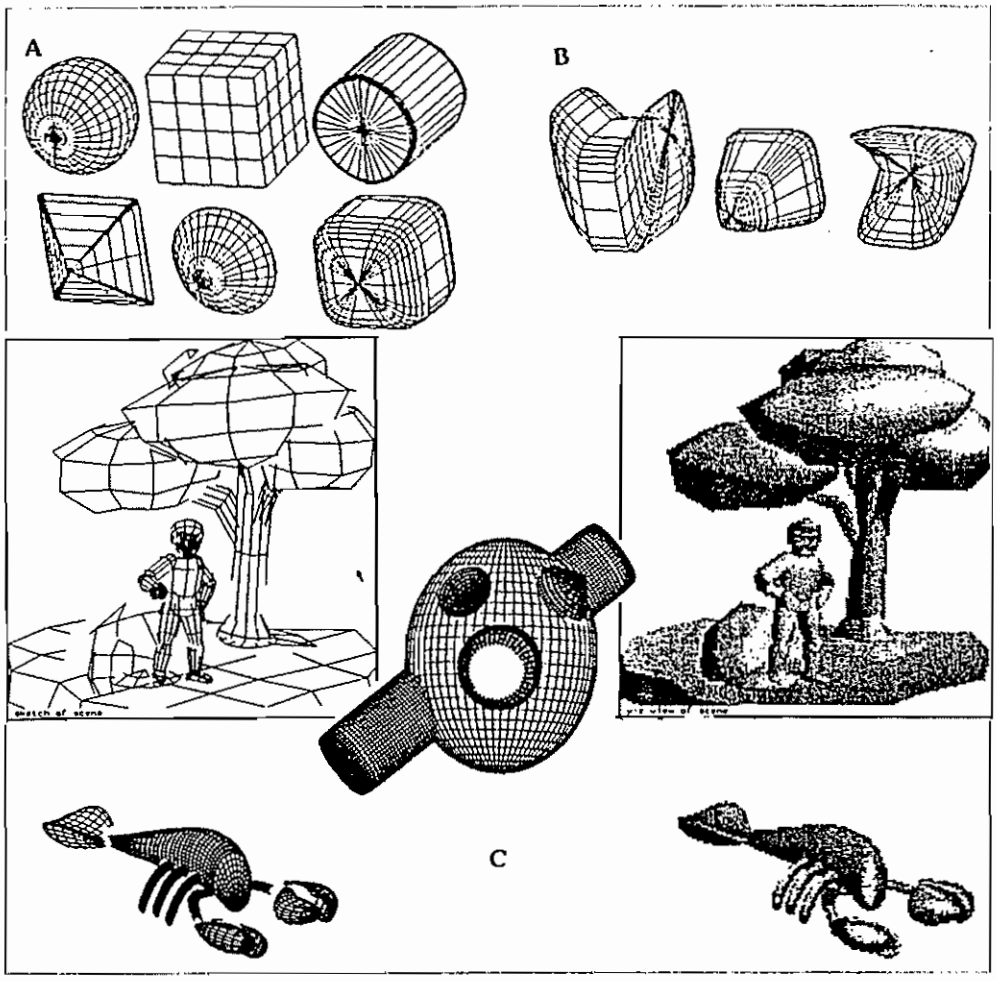



Fig. 3. (a) A sampling of the basic forms allowed. (b) Deformations of these forms. (c) Boolean combination (ors and nots) of the basic forms.

$$\boldsymbol{X}(\eta, \omega) = \begin{pmatrix} C_\eta^{\varepsilon_1} C_\omega^{\varepsilon_2} \\ c_\eta^{\varepsilon_1} S_\omega^{\varepsilon_2} \\ S_\eta^{\varepsilon_1} \end{pmatrix},$$

where $\boldsymbol{X}(\eta, \omega)$ is a three-dimensional vector that sweeps out a surface parameterized in latitude $\eta$ and longitude $\omega$, with the surface's shape controlled by the parameters $\varepsilon_1$ and $\varepsilon_2$. This family of functions includes cubes, cylinders, spheres, diamonds and pyramidal shapes as well as the round-edged shapes intermediate between these standard shapes. Some of these shapes are illustrated in Fig. 3(a). Superquadrics are, therefore, a superset of the modeling primitives currently in common use.

These basic "lumps of clay" (with various symmetries and profiles) are used as prototypes that are then deformed by stretching, bending, twisting or tapering, and then combined using Boolean operations to form new, complex prototypes that may, recursively, again be subjected to deformation and Boolean combination. As an example, the back of a chair is a rounded-edge cube that has been flattened along one axis, and then bent somewhat to accommodate the rounded human form. The bottom of the chair is a similar object, but rotated 90°, and by "oring" these two parts together with elongated rectangular primitives describing the chair legs we obtain a complete description of the chair, as illustrated in Fig. 4.

This descriptive language is designed to describe shapes in a manner that corresponds to a possible formative history, e.g., how one would create a given shape by combining lumps of clay. Thus the description provides us with an explanation of the image data in terms of the interaction of generic formative processes. This primitive explanation can then be refined by application of specific world knowledge and context, eventually deriving causal connections, affordances, and all of the other information that makes our perceptual experience appear so rich and varied. For instance, if we have parsed the chair in Fig. 4 into its constituent parts we could deduce that the bottom of the chair is a stable platform and thus might be useful as a seat, or we might hypothesize that the back of the chair can rigidly move relative to the supporting rod, given the evidence that they are separate "parts" and thus likely separately formed.

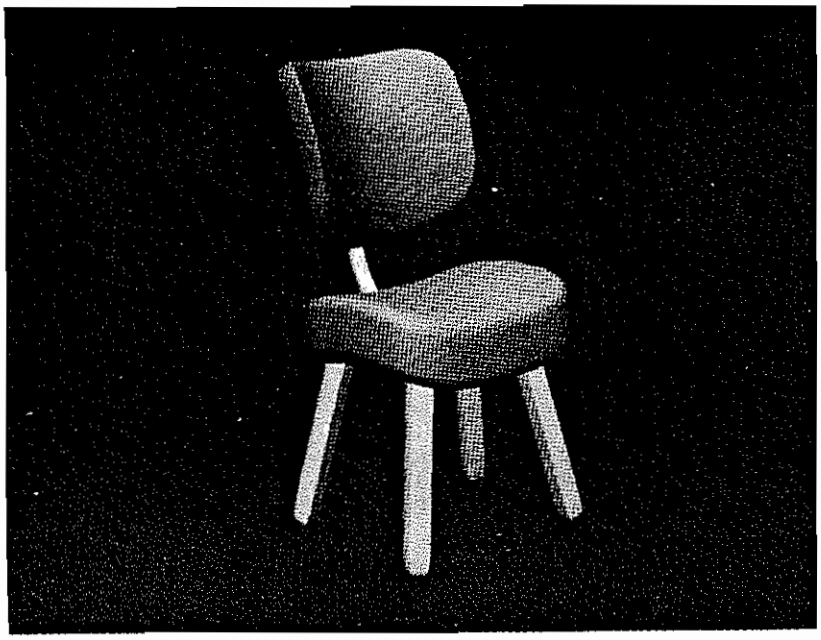


FIG. 4. A chair formed from Boolean combinations of appropriately deformed superquadrics.

The reader is encouraged to consider other examples of how the knowledge of part structure can help in forming hypotheses about function.

We have found that by using such a process-oriented, possible-history representation we force the resulting descriptions to group points that have similar causal histories, thus obtaining "parts" that interact with the world in a relatively simple, holistic manner. This further simplifies many reasoning tasks, because the parameters and components that affect interactions tend to be explicitly represented rather than being some complex or difficult-to-calculate function of the description's variables. For instance, use of this type of representation sufficiently simplfies questions about spatial relationships, intersection, image appearance, and so forth that we have been able to use it to construct a real-time 3-D graphical modeling system, using a Symbolics 3600 computer.[6] This system, called "SuperSketch," was used to make the figures in this paper.

Such descriptions may be written as a predicate calculus formula. We may then use this description, which has a clear model-theoretic semantics, in conjunction with constraint satisfaction or theorem-proving mechanisms, to accomplish whatever reasoning is required. Interestingly, it has been found that when adult human subjects are required to verbally describe imagery with completely novel content, their typical spontaneous strategy is to employ a descriptive system analogous to this one—i.e., form is described by modifying and combining prototypes [30]. The classic work by Rosch [3] supports the view that such a prototype-and-differences descriptive system is common in human reasoning: she showed that even primitive New Guinea tribesmen (who appear to have no concept of regular geometric shapes) form geometric prototypes in much the same manner as people from other cultures and describe novel shapes in terms of differences from these prototypes.

This representational system provides a grammar of form that has surprising descriptive power. Such descriptions have the intuitively satisfying nature of the Marr and Nishihara scheme; they incorporate hierarchies of primitives with axes of symmetry. This new descriptive language, however, is considerably more powerful than other representations that have been suggested. For example, a trivial comparison is that we can describe a wider range of basic shapes, as shown in Fig. 3(a). By allowing deformations of these shapes we greatly expand the range of primitives allowed, as shown in Fig. 3(b) (see also [31, 32, 42] on describing shape using modifications of prototypes). We have, so far, required only stretching, bending, tapering and twisting deformations to construct an extremely wide variety of objects. But the most powerful notion in

[6] Real-time" in this case means that a "lump" can be moved, hidden-surface removal accomplished, and drawn as a 100-polygon line-drawing approximation in one eighth of a second, and a complex (full-color) image such as Fig. 1 can be rendered in approximately 20 seconds. The Symbolics speed is roughly comparable to a VAX 11/780, except for being almost an order of magnitude slower on the floating-point operations that are used heavily in this modeling system.

this language is that of allowing (hierarchical) Boolean combination of these primitives. This intuitively attractive constructive solid-modeling approach—building specific object descriptions by applying the logical set operations "and," "or," and "not" to component *parts*—introduces a language-like generative power that allows the creation of a tremendous variety of form, such as is illustrated by Fig. 3(c) or by Fig. 1.

### 3.1. Biological forms

Biological forms such as the human body are naturally described by hierarchical Boolean combinations of the basic primitives, allowing the construction of accurate—but quite simple—descriptions of the detailed shape, as illustrated by Fig. 5 (the slightly cartoon-like nature of these illustrations is due primarily to the lack of surface texturing). The entire human body shown in Fig. 5, including face and hands, requires combining only 40 primitives, or approximately 300 bytes of information (these informational requirements are not a function of body position). Similarly, the description for the face requires the combination of only 13 primitives, or fewer than 100 bytes of information. The extreme brevity of these descriptions makes many otherwise difficult reasoning tasks relatively simple, e.g., even NP-complete problems can be easily solved when the size of the problem is small enough.

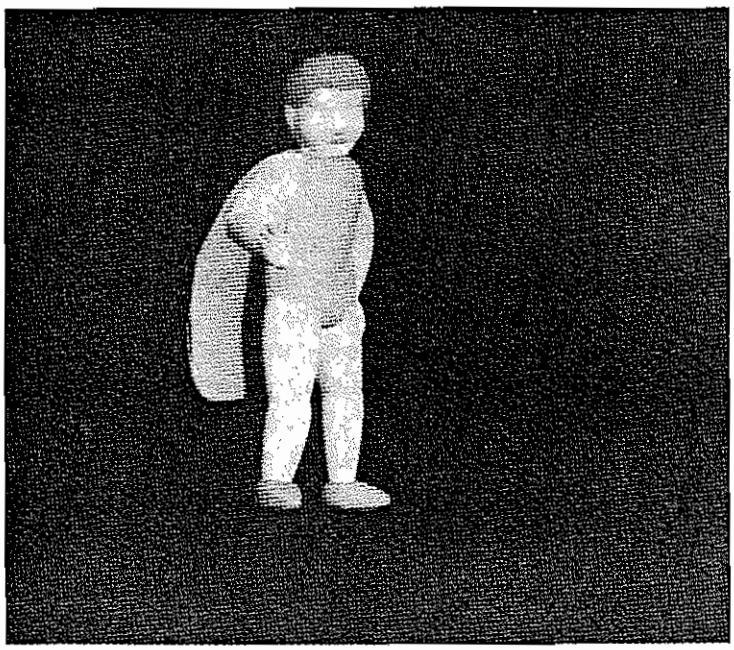


FIG. 5. The human form described (and rendered) by use of this representational system; only 40 primitives are required, approximately 300 bytes of information.

In Fig. 5 (as in all cases examined to date) when we try to model a particular 3-D form we find that we are able to describe—indeed, we are almost *forced* to describe—the shape in a manner that corresponds to the organization our perceptual apparatus imposes upon the image. That is, the components of the description match one-to-one with our naive perceptual notion of the "parts" in the figure, e.g., the face in Fig. 5 is composed of primitives that correspond exactly to the cheeks, chin, nose, forehead, ears, and so forth. Fig. 6 shows how the face is formed from the Boolean sum of several different primitives. The basic form for the head is a slightly tapered ellipsoid. To this basic form is

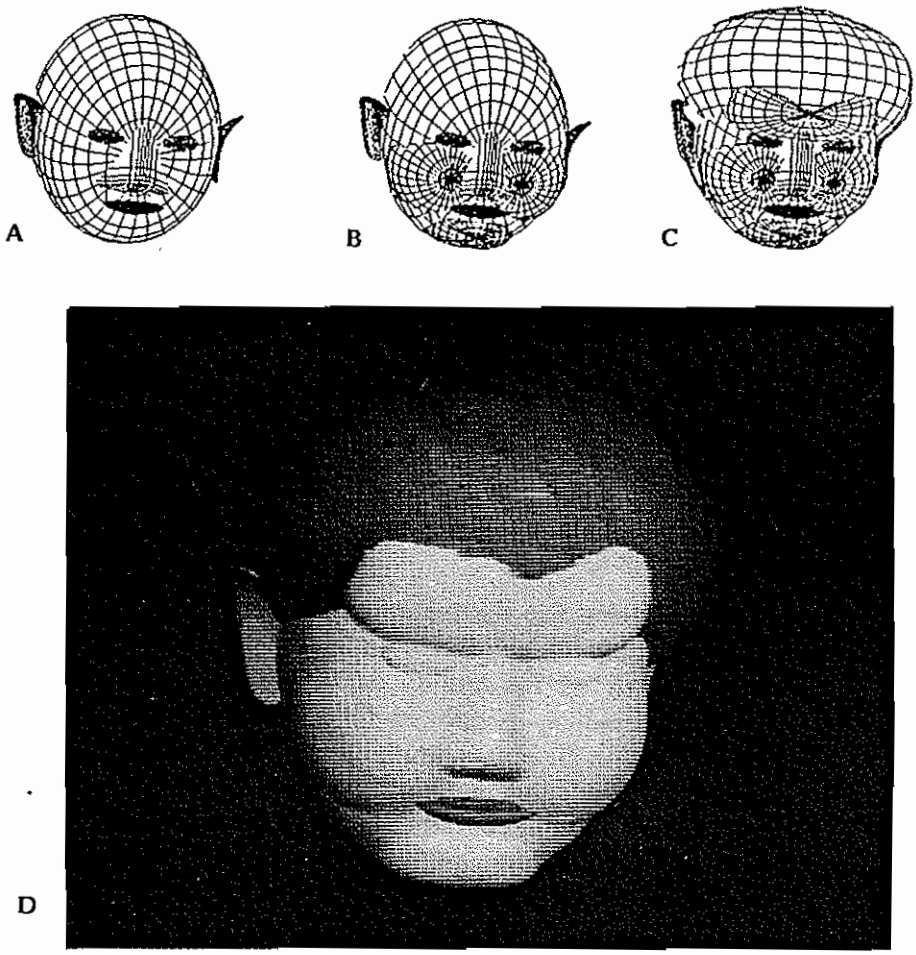



FIG. 6. (a) shows that the basic form for the head is a slightly tapered ellipsoid; to this basic form is added a somewhat cubical nose, bent pancake-like primitives for ears, bent thin ellipsoids for lips, and almond-shaped eyes. (b) shows the addition of rounded cheeks and slightly pointed chin (is this Yoda from Star Wars?), and finally (c) shows the addition of a squarish forehead and slightly fractalized hair. The smoothly shaded result is shown in (d)—it is a reasonably accurate human head, composed of only 13 primitives, specified by slightly less than 100 bytes of information.

added a somewhat cubical nose, bent pancake-like primitives for ears, bent thin ellipsoids for lips, and almond-shaped eyes, as is shown in Fig. 6(a). Fig. 6(b) shows the addition of rounded cheeks and a slightly pointed chin (is this Yoda from Star Wars?), and finally Fig. 6(c) shows the addition of a squarish forehead and slightly fractalized hair. The smoothly shaded result is shown in Fig. 6(d)—it is a reasonably accurate human head, composed of only 13 primitives, specified by slightly less than 100 bytes of information. One should remember that this representation is *not* in any way tailored for describing the human form: it is a general-purpose vocabulary.

The correspondence between the organization of descriptions made in this representation and human perceptual organization is important because it is strong evidence that we are on the right track. The fact that the distinctions made in this representation are very similar to those made by people makes it likely that descriptions couched in this language will be useful in understanding common-sense reasoning tasks, e.g., that the vocabulary of this representation might constitute a good set of primitive predicates for theories of common-sense reasoning such as sought by the Naive Physics [33] research program.[7]

Similarly, the ability to make the right "part" distinctions offers hope that we can form qualitative descriptions of specific objects ("Ted's face") or of classes of objects ("a long, thin face") by specifying constraints on part parameters and on relations between parts, in the manner of Marr and Nishihara [6,7], Winston [46, 47] or Davis [48]. And, of course, such representational correspondence is also important because it provides the basis for useful man–machine interaction.

### 3.2. Complex inanimate forms

This method for representing the three-dimensional world, although excellent for biological and man-made forms, becomes awkward when applied to complex natural surfaces such as mountains or clouds. The most pronounced difficulty is that, like previously proposed representations, our superquadric lumps-of-clay representation becomes implausably complex when confronted with the problem of representing, e.g., a mountain, crumpled newspaper, a bush or a field of grass. This makes the technique ill-suited to solving the problem of representing *classes* of such objects, or determining that a particular object is a member of that class.

Why is it that such introspectively simple tasks turn out to be so hard? Intuitively, the main source of difficulty is that there is too much information to

[7] Descriptions that correspond to a possible formative history explicitly group together parts of a form that have a similar causal history, i.e., that came about in the same manner. It appears that such groupings have a strong tendency to *continue* to act as a simple whole. Why this should be true is unclear; perhaps there are only a few basic categories of physical interaction that all may be characterized using the same definition of "part."

deal with. Natural objects are amazingly bumpy and detailed; and classes of such objects seem to include virtually infinite variability. There is simply too much detail, and it is too variable. When we attempt to represent such objects in a detailed, quantitative manner, we are forced to an unwieldy description.

Nor does it suffice to simply introduce error tolerances into the representation, e.g., a mountain is a cone $\pm x$. For not only is such a representation misleading (do we *really* want to say that a cube is a sphere $\pm 0.25r$?), but it does not allow for the ability to distinguish between a mountain (represented as a cone $\pm x$) and a cone with a few dents in it (also represented as a cone $\pm x$).

Experiments in human perception suggest a way out of such problems. When we view a crumpled newspaper (for instance), it seems that the description we store is not accurate enough to recover every detail; rather, it seems that out of the welter of image detail people abstract a few properties such as the general "crumpledness" and a few major features of the shape, e.g., the general outline. The rest of the crumpled newspaper's structure is ignored; it is unimportant, *random*. For the purpose of describing that crumpled newspaper, then, the only important constraints on shape are the crumpledness and general outline.

People escape the trap of overwhelming complexity, it seems, by varying the level of descriptive abstraction—the amount of detail captured—depending on the task. In cases like the crumpled newspaper, or when recognizing classes of objects such as "a mountain" or "a cloud," the level of abstraction is very high. Almost no specific detail is required, only that the crumpledness of the form comply with the general physical properties characteristic of that type of object. In recognizing a specific mountain, however, people will require that all of the major features be identical, although they typically ignore smaller details. Even though these details are "ignored," however, they must still conform to the constraints characteristic of that type of object: we would never mistake a smooth cone for a rough-surfaced mountain even if it had a generally conical shape.

The fractal model of natural surfaces [34, 35] allows us to duplicate this sort of physically meaningful abstraction from the morass of details encountered in natural scenes. It lets us describe a crumpled newspaper by specifying certain structural regularities—its crumpledness, in effect—and leave the rest as variable detail. It lets us specify the qualitative shape—i.e., the surface's roughness—without (necessarily) worrying about the details.

### 3.2.1. *Fractal-based qualitative description*

Many naturally occurring forms are fractals[8] [22, 34–36]; Mandelbrot, for

[8] The defining characteristic of a fractal is that it has a *fractional dimension*, from which we get the word "fractal."

instance, shows that fractal surfaces are produced by several basic physical processes. One general characterization of naturally occurring fractals is that they are the end result of any physical processes that randomly modifies shape through local action, i.e., they are a generalization of random walks and Brownian motion. After innumerable repetitions, such processes will typically produce a fractal surface shape. Thus clouds, mountains, turbulent water, lightning and even music have all been shown to have a fractal form.

During the last two years we have developed these fractal functions into a statistical model for describing complex, natural surface shapes [34, 35, 37], and have found that it furnishes a good description for many surfaces. Evidence for the descriptive adequacy of this model comes from several sources. Recently conducted surveys of natural imagery [34–36], for instance, have found that this model accurately describes how most homogeneous textured or shaded image regions change over scale (change in resolution). The prevalence of surfaces with fractal statistics is explained by analogy to Brownian motion (the archetypical fractal function): just as when a dust mote randomly bombarded by air molecules produces a fractal Brownian random walk, the complex interaction of processes that locally modify shape produces a fractal Brownian surface.

For our current purposes, perhaps the most important fact is that one of the parameters of this statistical model (specifically, the fractal dimension of the surface) has been found to correspond very closely to people's perceptual notion of *roughness* [38, 39]. We have been able, for instance, to accurately predict a surface's perceptual smoothness or roughness on the basis of knowing it's fractal statistics. The fractal model, therefore, gives us a way of *qualitatively* describing surface shape [34, 35].

The fractal model shows how we may use physically motivated statistical description to abstract away from the overwhelming amount of detail present in many natural forms. To be useful, however, we must combine the fractal model's notion of qualitative description by physically meaningful statistical abstraction together with the quantitative descriptive abilities of the lump-of-clay descriptive language developed in the previous sections.

### 3.2.2. *Qualitative and quantitative description*

We begin the task of unifying the fractal model's notion of qualitative description with the quantitative lump-of-clay description by considering the basic properties of naturally occurring examples of fractal Brownian surfaces. Such surfaces all have two important properties: (1) each segment is statistically similar to all others; (2) segments at different scales are statistically indistinguishable, i.e., as we examine such a surface at greater or lesser imaging resolution its statistics (curvature, etc.) remain the same. Because of these invariances, the most important *variable* in the description of such a shape is how it varies with scale; in essence, how many large features there are relative to the number of middle-sized and smaller-sized features. For fractal shapes (and thus for many real shapes) the ratio of the number of features of one size to

the number of features of the next larger size is a constant—a surprising fact that derives from the property of scale invariance. The fractal model, therefore, leads us to a statistical characterization of a surface in terms of two parameters: the surface's variance (amplitude), and the ratio between the frequency of smaller and larger features (i.e., its fractal dimension).

We may, therefore, construct fractal surfaces by using our superquadric "lumps" to describe the surface's features; specifically, we can use the recursive sum of smaller and smaller superquadric lumps to form a true fractal surface. This construction is illustrated in Figs. 7(a)–(c).

We start by specifying the surface's qualitative appearance—its roughness—by picking a ratio $r$, $0 \leq r \leq 1$, between the number of features of one size to the number of features that are twice as large. This ratio describes how the surface varies across different scales (resolutions, spatial frequency channels, etc.) and is related to the surface's fractal dimension $D$ by $D = T + r$, where $T$ is the topoligical dimension of the surface.

We then randomly place $n^2$ large bumps on a plane, giving the bumps a Gaussian distribution of altitude (with variance $\sigma^2$), as seen in Fig. 7(a). We then add to that $4n^2$ bumps of half the size, and altitude variance $\sigma^2 r^2$, as shown in Fig. 7(b). We continue with $16n^2$ bumps of one quarter the size, and altitude $\sigma^2 r^4$, then $64n^2$ bumps one-eighth size, and altitude $\sigma^2 r^6$ and so forth, as shown in Fig. 7(c). The final result, shown in Fig. 7(c) is a true Brownian fractal shape. The validity of this construction does not depend on the particular shape of the superquadric primitives employed, the only constraint is that the sum must fill out the Fourier domain. Different shaped lumps will, however, give different appearance or texture to the resulting fractal surface; this is an important and as yet relatively uninvestigated aspect of the fractal model. Figs. 7(d) and 7(e) illustrate the power and generality of this construction; all of the forms and surfaces in these images can be constructed in this manner.

When the placement and size of these superquadric lumps is random, we obtain the classical Brownian fractal surface that has been the subject of our previous research. When the larger components of this sum are matched to a particular object, however, we obtain a description of that object that is exact to the level of detail encompassed by the specified components. This makes it possible to specify a global shape while retaining a qualitative, statistical description at smaller scales: to describe a complex natural form such as a cloud or mountain, we specify the "lumps" down to the desired level of detail by fixing the larger elements of this sum, and then we specify only the fractal statistics of the smaller lumps thus fixing the qualitative appearance of the surface. Fig. 8 illustrates an example of such description. The overall shape is that of a sphere; to this specified large-scale shape, smaller lumps were added randomly. The smaller lumps were added with six different choices of $r$ (i.e., six different choices of fractal statistics) resulting in six qualitatively different surfaces—each with the same basic spherical shape.

The ability to fix particular "lumps" within a given shape provides an elegant

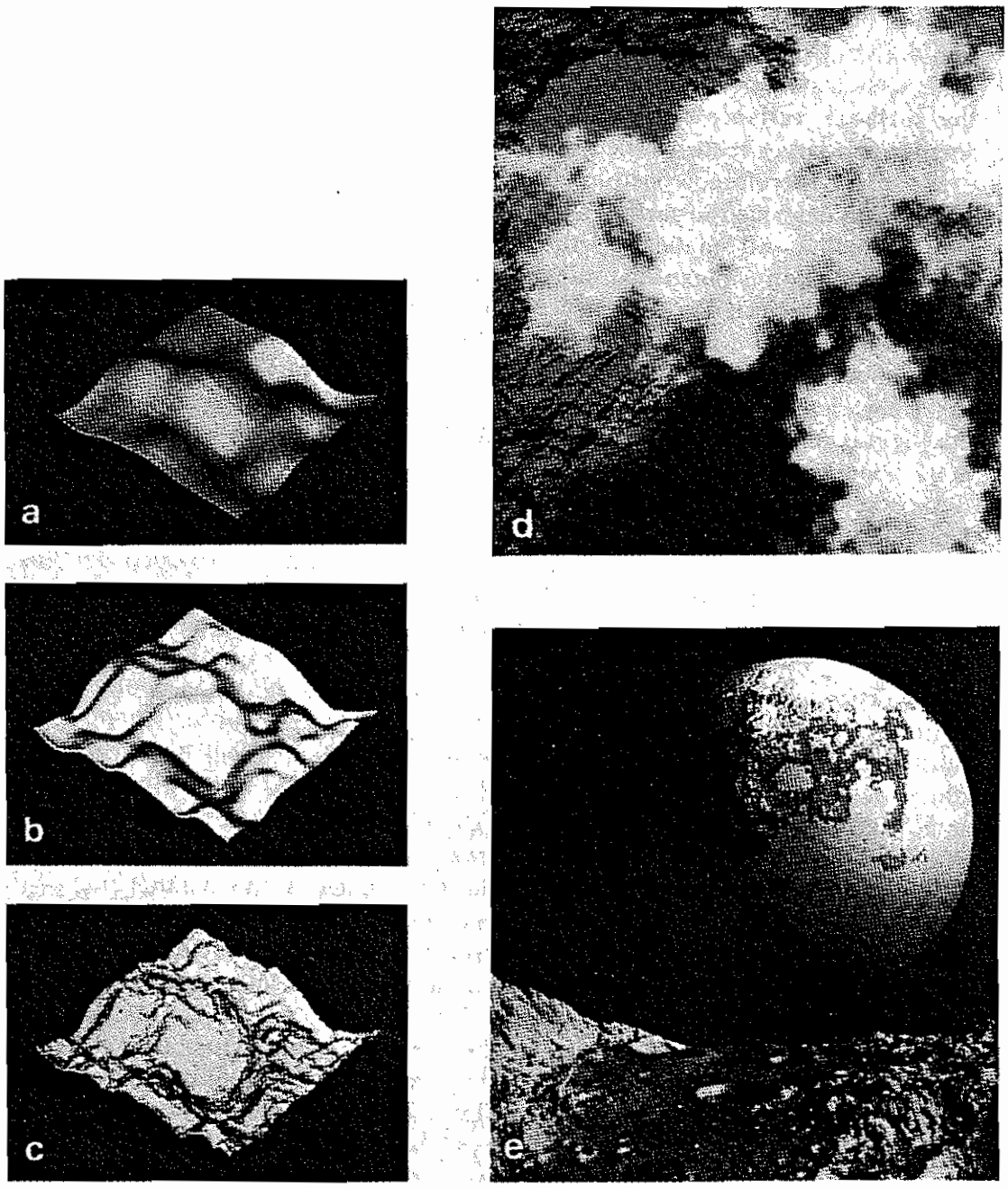



FIG. 7. (a)–(c) show the construction of a fractal shape by successive addition of smaller and smaller features with number of features and amplitudes described by the ratio $1/r$. All of the forms and surfaces shown in (d) and (e) (which are images by Voss and Mandelbrot, see [22]) can be generated in this manner.

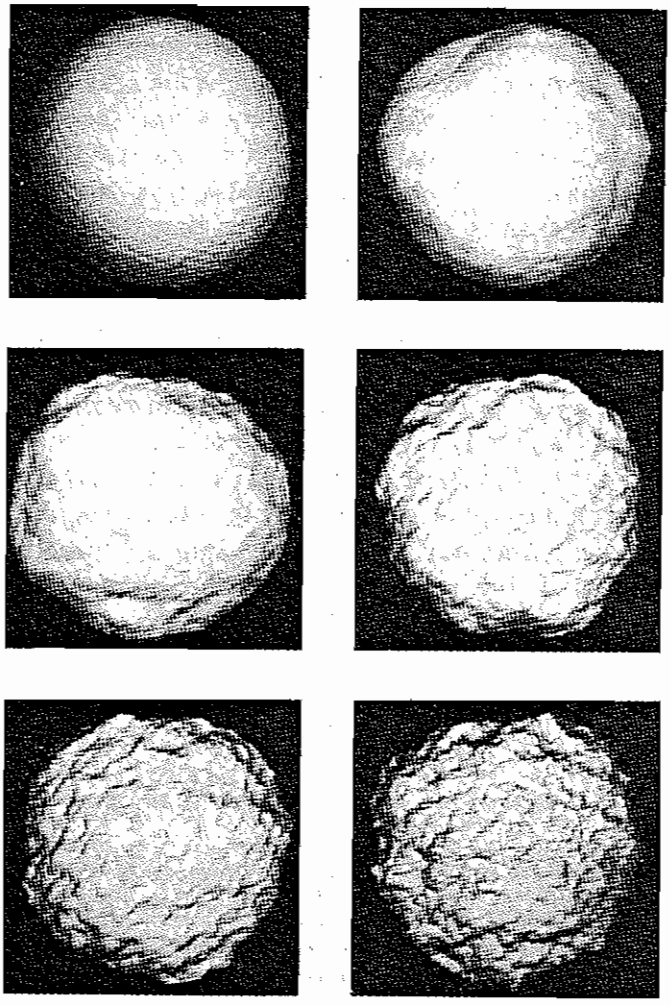


FIG. 8. Spherical shapes with surface crenulations ranging from smooth (fractal dimension = topological dimension, $r \approx 0$) to rough (fractal dimension $\gg$ topological dimension, $r \approx 1$).

way to pass from a qualitative model of a surface to a quantitative one—or *vice versa*. We can refine a general model of the class "a mountain" to produce a model of a *particular* mountain by fixing the position and size of the largest lumps used to build the surface, while still leaving smaller details only statistically specified. Or we can take a very specific model of a shape, discard the smaller constituent lumps after calculating their statistics, and obtain a model that is less detailed than the original but which is still *qualitatively* correct.

## 4. Primitive Perception: Recognizing Instances of Models

During the last decade, the dominant view of human perception has been that perception proceeds through successive levels of increasingly sophisticated

representations until finally, at some point, information is transferred to our general cognitive faculties. And indeed, there *does* seem to be a gradient of sophistication in human perception, ranging from seemingly primitive inferences of shapes, textures, colors, and the like, to the apparently more sophisticated inferences of chairs, trees, affordances[9] and people's emotions. There is significant reason to believe, however, that this is not simply the flow of information through successive levels of representation.

To summarize Fodor's excellent and extended argument for this conclusion [40], we note that the sophisticated end of perception can involve virtually anything we know, and seems to blend smoothly into general cognition—for instance, we speak of perceiving abstract mathematical relationships or people's intentions. There is no principled reason to separate sophisticated perception from general-purpose reasoning. The characteristics of primitive perception, however, are quite different from that of cognition:

- *Informational encapsulation.* Primitive perception proceeds without benefit of intimate access to the full range of our world knowledge. Most visual illusions, for instance, cannot be dispelled merely by recognizing them as illusions [41].
- *Limited extent.* The body of knowledge on which primitive perception draws is of quite limited extent, at least in comparison to our conscious world knowledge. People of all cultures seem to share a common perceptual framework [43]; it is this shared framework that makes possible any communication at all.
- *Functional autonomy.* Primitive perception proceeds with little regard to the particulars of the task at hand, under at most limited voluntary control. We are capable of the same discriminations, regardless of purpose or task, except (perhaps) for a few very practiced tasks, e.g., birdwatchers discriminating between different types of bird. This is not to say that we always *do* make the same discriminations (we can, after all, focus our attention), but rather that whenever we attend to a particular stimulus dimension we are always capable of making the same discriminations.

Primitive perception is at least roughly the realm of perceptual organization, i.e., the pre-attentive organization of sensory data into primitives like texture, color and form. Thus, although we often speak as if perception were a smooth series of progressively more sophisticated inferences (e.g., Marr [10]), it is more likely that there are separate, specialized mechanisms for primitive and sophisticated perception.

This leads to a conception of our perceptual apparatus as containing two distinct parts: the first, a special-purpose, perhaps innate mechanism that supports primitive perception, and the second something that closely resembles general cognition. Most of the time the sensory data are first examined by the mechanisms of primitive perception to discover instances of rigidity,

[9] Affordances are the purpose(s) of an object.

parallelism, part-like groupings and other evidences of causal organization, thus providing an explanation of the image data in terms of generic formative processes. The mechanisms of sophisticated perception then use specific, learned knowledge about the world to refine this primitive, generic explanation into a detailed account of the environment.

It should be noted, however, that for at least the most practiced discriminations things seem to happen somewhat differently. When a percept, even if of a very sophisticated nature, is highly practiced or very important it appears that our minds build up a special-purpose mechanism solely for that purpose. Consider, for instance, our incredible facility at recognizing our own name, or the faces of familiar people. There may be, therefore, a sort of "compiler" for building specialized routines for these oft-repeated, important or time-critical discriminations. How much of our day-to-day perception is handled by such special-purpose routines is very much an open question.

Primitive perception, by our definition, was first seriously addressed by the Gestalt psychologists [4], who noticed that people seem to spontaneously impose a physically meaningful organization upon visual stimuli, through grouping, figure/ground separation, and so forth. They found that the addition of semantic context very rarely affects this spontaneous, pre-attentive organization of the image; somehow the visual system seems able to group an image into the correct, physically meaningful parts *before* contextual knowledge is available.

The Gestalt psychologists described this spontaneous organization as being governed by the principle of *Pragnanz*[10], however their lack of modern notions of computation limited their ability to crisply define *Pragnanz* and thus doomed them to a rather limited success. Nevertheless, their work paved the way for the two-stage model of perception that is enjoying widespread popularity in academic circles today. The first stage, which we are describing here as primitive perception, is spontaneous and pre-attentive. It carves the sensory data into likely-meaningful parts, and presents them to the later stages of perception. The second stage of perception, which we are calling sophisticated perception, is very little (if at all) different from our general cognitive faculty—including the ability to make very efficient, "compiled" routines, presumably by combining the outputs of primitive perception.

### 4.1. Recognizing our modeling primitives

It is our goal to provide the beginnings of a theory for our faculty of pre-attentive, primitive perception: to present a rigorous mathematical definition for the vague notion of "a part" and to explain how we can, Gestalt-like, carve an image up into meaningful "parts" without need of semantic context or specific a priori knowledge. We have already described a representation that is competent to describe a wide range of natural forms, and

[10]Pragnanz is normally translated as meaning "goodness of form."

whose primitive elements seem to correspond closely to our naive notions of perceptual parts. What remains to be done is to show that these descriptive primitives can be recovered from the image data.

The major difficulty in recovering such descriptions is that image data are mostly a function of surface normals, and not directly a function of the surface shape. This is because image intensity, texture anisotropy, contour shape, and the like—the information we have about surface shape—is largely determined by the direction of the surface normal. To recover the shape of a general volumetric primitive, therefore, we must (typically) first compute a dense depth map from information about the surface normals. The computation of such a depth map has been the major focus of effort in vision research over the last decade and, although the final results are not in, the betting is that such depth maps are impossible to obtain in the general, unconstrained situation. Even given such a depth map, the recovery of a shape description has proven extremely difficult, because the parameterization of the surface given in the depth map is generally unrelated to that of the desired description.

Because image information is largely a function of the surface normal, one of the most important properties of superquadrics is the simple "dual" relation between their surface normal and their surface shape. It appears that this dual relationship can allow us to form an overconstrained estimate of the 3-D parameters of such a shape from noisy or partial image data, as outlined by the following equations.

The surface position vector of a superquadric with length, width and breadth $a_1$, $a_2$ and $a_3$ is (again writing $\cos\eta = C_\eta$, $\sin\omega = S_\omega$)

$$\boldsymbol{X}(\eta, \omega) = \begin{pmatrix} a_1 C_\eta^{\epsilon_1} C_\omega^{\epsilon_2} \\ a_2 C_\eta^{\epsilon_1} S_\omega^{\epsilon_2} \\ a_3 S_\eta^{\epsilon_1} \end{pmatrix} \tag{1}$$

and the surface normal at that point is

$$\boldsymbol{N}(\eta, \omega) = \begin{pmatrix} \frac{1}{a_2} C_\eta^{2-\epsilon_1} C_\omega^{2-\epsilon_2} \\ \frac{1}{a_2} C_\eta^{2-\epsilon_1} S_\omega^{2-\epsilon_2} \\ \frac{1}{a_3} S_\eta^{2-\epsilon_1} \end{pmatrix}. \tag{2}$$

Therefore the surface vector $\boldsymbol{X} = (x, y, z)$ is dual to the surface normal vector $\boldsymbol{N} = (x_n, y_n, z_n)$ in the following sense:

$$\boldsymbol{N}(\eta, \omega) = \begin{pmatrix} \frac{1}{x} C_\eta^2 C_\omega^2 \\ \frac{1}{y} C_\eta^2 S_\omega^2 \\ \frac{1}{z} S_\eta^2 \end{pmatrix}. \tag{3}$$

From (1) and (3), then we have

$$x_n = \frac{C_\eta^2 C_\omega^2}{x}, \tag{4}$$

$$y_n = \frac{C_\eta^2 S_\omega^2}{y}, \tag{5}$$

so

$$\frac{y_n}{x_n} = \frac{x}{y} \tan^2 \omega . \tag{6}$$

or

$$\left(\frac{y y_n}{x x_n}\right)^{1/2} = \tan \omega . \tag{7}$$

We may also derive alternative expressions for tan $\omega$ as follows:

$$x = a_1 C_\eta^{r_1} C_\omega^{r_2}, \qquad y = a_2 C_\eta^{r_1} S_\omega^{r_2}, \tag{8}$$

so

$$\frac{x}{y} = \frac{a_1}{a_2} \left(\frac{C_\omega}{S_\omega}\right)^{\varepsilon_2} \tag{9}$$

or

$$\left(\frac{y a_1}{x a_2}\right)^{1/\varepsilon_2} = \tan \omega . \tag{10}$$

Combining these expressions for tan $\omega$ we obtain

$$\left(\frac{y y_n}{x x_n}\right)^{1/2} = \left(\frac{y a_1}{x a_2}\right)^{1/\varepsilon_2} \tag{11}$$

or

$$\frac{y_n}{x_n} = \left(\frac{y}{x}\right)^{2/\varepsilon_2 - 1} \left(\frac{a_1}{a_2}\right)^{2/\varepsilon_2} . \tag{12}$$

Letting $\tau = y_n/x_n$, $k = (a_1/a_2)^{2/\varepsilon_2}$ and $\xi = 2/\varepsilon_2 - 1$ we find that

$$\tau = k \left(\frac{y}{x}\right)^{\xi}, \tag{13}$$

$$\frac{\mathrm{d}\tau}{\mathrm{d}y} = \frac{k\xi}{x} \left(\frac{y}{x}\right)^{\xi - 1}, \tag{14}$$

$$\frac{d\tau}{dx} = \frac{-k\xi y}{x^2}\left(\frac{y}{x}\right)^{\epsilon-1}. \tag{15}$$

This gives us two equations relating the unknown shape parameters to image measurable quantities, i.e.,

$$\frac{\tau}{d\tau/dy} = \frac{y}{\xi} \tag{16}$$

and

$$\frac{\tau}{d\tau/dx} = \frac{-x}{\xi}. \tag{17}$$

Thus (16) and (17) allow us to construct a linear regression to solve for center and orientation of the form, as well as the shape parameter $\epsilon_2$, given only that we can estimate the surface-tilt direction $\tau$.

#### 4.1.1. *Overconstraint and reliability*

Perhaps the most important aspect of these equations is that we can form an *overconstrained* estimate of the 3-D parameters, we can *check* that the parameters we estimate are correct. This property of overconstraint comes from using models: when we have used some points on a surface to estimate 3-D parameters, we can check if we are correct by examining additional points. The model predicts what these new points should look like; if they match the predictions, then we can be sure that the model applies and that the parameters are correctly estimated. If the predictions do *not* match the new data points, then we know that something is wrong. The ability to check your answer is perhaps the most important property any vision system can have, because only when you can check your answers can you build a reliable vision system. And it is *only* when you have a model that relates many different image points (such as a model of how rigid motion appears in an image sequence, or a CAD-CAM model, or this 3-D shape model) that you can have the overconstraint needed to check your answer.

One other aspect of (16) and (17) that deserves special note is that the only image measurement needed to recover 3-D shape is the surface tilt $\tau$, the component of shape that is unaffected by projection and, thus, is the most reliably estimated parameter of surface shape. It is, for instance, known exactly at smooth occluding contours and both shape-from-shading and shape-from-texture methods produce a more reliable estimate of $\tau$ than of slant, the other surface shape parameter. That we need only the (relatively) easily estimated tilt to estimate the 3-D shape parameters makes robust recovery of 3-D shape much more likely.

When we generalize these equations to include unknown orientation and position parameters for the superquadric shape, we obtain a new set of nonlinear equations that can then be solved (in closed form) for the unknown shape parameters $\varepsilon_1$ and $\varepsilon_2$, the center position, and the three angles giving the objects orientation. Once these unknowns are obtained the remaining unknowns ($a_1$, $a_2$, and $a_3$, the three dimensions of the object) are easily obtained.

For the case of rotation and translation in the image plane, the equations become:

$$x^* = C_\theta(x - x_0) + S_\theta(y - y_0), \qquad y^* = -S_\theta(x - x_0) + C_\theta(y - y_0) \tag{18}$$

where $\theta$ is the rotation, $x_0$, $y_0$ the translation, and $(x^*, y^*)$ the new rotated and translated coordinate system. The tilt $\tau$ then becomes

$$\tau = \frac{y_n^*}{x_n^*} = \frac{(-S_\theta x_n + C_\theta y_n)}{(C_\theta x_n + S_\theta y_n)}, \tag{19}$$

and the derivative of (19) is

$$\begin{aligned}\frac{d\tau}{dy^*} &= \left(-S_\theta \frac{dx_n}{dy^*} + C_\theta \frac{dy_n}{dy^*}\right)(C_\theta x_n + S_\theta y_n)^{-1} \\ &\quad -(-S_\theta x_n + C_\theta y_n)(C_\theta x_n + S_\theta y_n)^{-2}\left(C_\theta \frac{dx_n}{dy^*} + S_\theta \frac{dy_n}{dy^*}\right) \\ &= (C_\theta x_n + S_\theta y_n)^{-2}\left(x_n \frac{dy_n}{dy^*} - y_n \frac{dx_n}{dy^*}\right).\end{aligned} \tag{20}$$

Noting that

$$\begin{aligned}\frac{dx_n}{dy^*} &= \frac{dx_n}{dx}\frac{dx}{dy^*} + \frac{dx_n}{dy}\frac{dy}{dy^*} = -\frac{dx_n}{dx} S + \frac{dx_n}{dy} C, \\ \frac{dy_n}{dy^*} &= \frac{dy_n}{dx}\frac{dx}{dy^*} + \frac{dy_n}{dy}\frac{dy}{dy^*} = -\frac{dy_n}{dx} S + \frac{dy_n}{dy} C.\end{aligned} \tag{21}$$

Equation (16) can now be rewritten as

$$\begin{aligned}&(C_\theta x_n + S_\theta y_n)(-S_\theta x_n + C_\theta y_n) \\ &\quad = \frac{1}{\xi}\left[-S_\theta(x - x_0) + C_\theta(y - y_0)\right] \\ &\qquad \times \left(x_n\left(-\frac{dy_n}{dx} S + \frac{dy_n}{dy} C\right) - y_n\left(-\frac{dx_n}{dx} S + \frac{dx_n}{dy} C\right)\right).\end{aligned} \tag{22}$$

Our estimates of tilt from local image information typically have considerable noise in them [18, 37, 44]; in order to still obtain good estimate of three-dimensional shape we will formulate the problem of recovering the shape parameters as a linear regression. Collecting the image-measurable terms together (in square brackets), this equation becomes

$$\begin{aligned} 0 = {} & [x_n^2 - y_n^2](\xi C_\theta S_\theta) + [x_n y_n](\xi(S_\theta^2 - C_\theta^2)) \\ & + \left[ x x_n \frac{dy_n}{dy} - x y_n \frac{dx_n}{dy} \right](-S_\theta C_\theta) + \left[ x x_n \frac{dy_n}{dx} - x y_n \frac{dx_n}{dx} \right](S_\theta^2) \\ & + \left[ y x_n \frac{dy_n}{dy} - y y_n \frac{dx_n}{dy} \right](C_\theta^2) + \left[ y x_n \frac{dy_n}{dx} - y y_n \frac{dx_n}{dx} \right](-S_\theta C_\theta) \\ & + \left[ x_n \frac{dy_n}{dy} - y_n \frac{dx_n}{dy} \right](S_\theta C_\theta x_0 - C_\theta^2 y_0) \\ & + \left[ x_n \frac{dy_n}{dx} - y_n \frac{dx_n}{dx} \right](S_\theta C_\theta y_0 - S_\theta^2 x_0) . \end{aligned} \tag{23}$$

This equation, then, can be used for a linear regression to solve for the unknown coefficients (in curved brackets). We have seven unknown coefficients and so we require tilt information at as few as seven points in order to solve for all these unknowns. By combining this equation together with the equivalent version derived from (17) we can obtain closed form solutions for the center of the object $(x_0, y_0)$, the shape parameter $\varepsilon$, and the orientation $\theta$. In fact, things are somewhat better than this, because we have two such equations at each point (one for $dx$ and one for $dy$) so that fewer points are actually required. The small number of points required opens up the possibility of segmenting images in terms of the parameters of the 3-D surface.

At occluding contours the situation is better yet, because we also know that $y_n^2 + x_n^2 = 1$, and considerable extra constraint is available. This formulation, therefore, reflects the fact that contour information is more powerful than shading or texture information. One of the more interesting aspects of this approach is that contour information and information from shading or texture contribute toward estimating shape in exactly the same manner—by providing information about surface tilt—and therefore we may combine information from all of these sources by use of the same set of equations, those derived from (16) and (17).

Because we have formulated the problem of primitive perception as one of recognizing instances of the "parts" found in a representational vocabulary, we can frame the problem as one in statistical decision theory: we have a range of alternatives that we entertain, and use image data to decide among the alternatives. This gives us a rigorous framework for integrating information from motion, stereo, etc., together with contour, shading and texture infor-

mation without having to make further assumptions. This is in considerable contrast to approaches that try to apply strong, unverifiable assumptions about the nature of surfaces (e.g., that all surfaces are "smooth") in order to integrate various information sources. Here we are attempting to collect a vocabulary of models that span the space of shape possibilities, so that we can replace unverifiable assumptions with verifiable models. We want perception to proceed by making an overconstrained, statistical determination that a particular model is applicable (rather than simply making an assumption), and then estimate the parameters of that model. If our vocabulary of shape does in fact

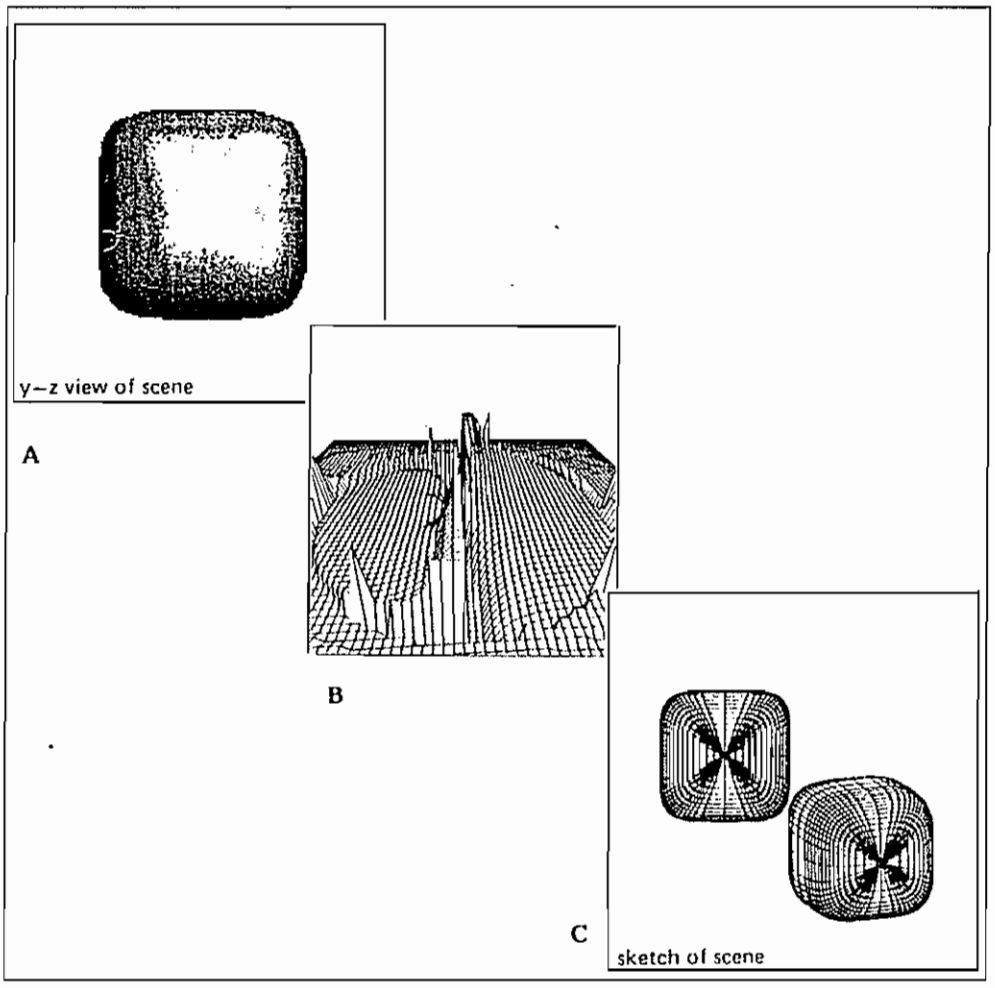



FIG. 9. (a) A half-toned version of an image of a superquadric with shape parameters $\varepsilon_1, \varepsilon_2 = 0.5$. (b) The surface tilts estimated using the local shape-from-shading/texture algorithm described in [18, 37]. (c) Two views of the 3-D shape estimated by use of equations (16) and (17), using the tilt estimates shown in (b).

cover the range of shape that actually occurs, then *we will have made the best shape estimate possible with the available image data.*

Equation (23) does not reflect the full sophistication possible in statistical decision theory; a regression using this equation results in a maximum likelihood estimate of compound parameters such as $C_\theta$ and $\xi C_\theta S_\theta$ rather than estimates of the individual parameters $\varepsilon$ and $\theta$. Still, the main power of the approach remains. Our modeling primitives provide us with a parameterized range of hypotheses that we can choose among using established statistical tools, thus providing us a rigorous framework for integrating contour with shading and texture information, as well as allowing us to include a priori information that we may have gained from previous views. The power of this framework has been illustrated by the work of Ferrie and Levine [44] who, using a simpler shape vocabulary consisting of ellipsoids and cylinders, have combined our local shape-from-shading technique [18] with motion information to accurately recover 3-D shape.

Although the equations presented here are only for rotations in the image plane, the general equations are similar, although somewhat more complicated. As in the simpler case, information at relatively few points[11] is required in order to solve for the unknowns, and the situation is considerably better along occluding contours.

Fig. 9 illustrates the process of recovering 3-D shape using this technique. Fig. 9(a) shows a half-toned version of an image of a superquadric with shape parameters $\varepsilon_1, \varepsilon_2 = 0.5$. To this image, we applied the local shape-from-shading/texture technique developed by Pentland [18, 37]. The estimation technique employs second-derivative filters with local support to make estimates of surface slant and tilt, with the estimates of tilt being more reliable than the estimates of slant [18, 44]. Fig. 9(b) shows a view of the surface tilt (i.e., $y_n/x_n$) recovered from the continuous 8-bit image of the shape illustrated by Fig. 9(a); in this figure the image $x$-axis runs left-right and the $y$-axis runs up-down. From this estimated tilt surface we can use (16) and (17) to estimate the center of the shape, the shape parameter $\varepsilon_2$, and the width and breadth of the shape. Fig. 9(c) shows two views of the recovered shape; it can be seen that in this simple case a good estimate of the 3-D shape can be made. It appears, then, that (16) and (17) offer a good hope for recovering surface shape; in our future research we hope to extend these preliminary results to natural imagery.

### 4.2. Model-based vision, the blocks world, and our effort

The most successful (i.e., working, practical) efforts in machine vision have all been accomplished within two paradigms that are generally lumped together under the rubric of "model-based vision." The first of these paradigms is to

[11] Depending on the exact formulation, 15 points are required.

take a CAD-CAM-type model of a specific object, find configurations of image features that uniquely determine identity and orientation of the object, and then search the image for those configurations. A similar, but fundamentally quite different "model-based vision" paradigm was first employed in the blocks-world research during the 1960s [45], and more recently in such work as the 3-D mosaic program of Hermann and Kanade [24]. In this second paradigm, the models are of the *parts* that make up specific objects, rather than a model of the entire object, and the goal is to identify those component parts. Once the parts have been identified and their spatial layout determined, one can ask if this configuration of parts is an object that has been seen before. This latter approach has the very significant advantage that it can learn new object descriptions by example: it can look at a new object, identify the object's parts, and then use that part-wise description to build up a general model of the specific object in a manner similar to that proposed by Winston [46, 47].

Because this second, find-the-parts approach to model-based vision can learn descriptions of novel objects, it has the potential to support general-purpose vision. The major limitation on the success of this approach is the availability of part models that are individually recognizable and which have the expressive power to describe everything within the domain of interest. What we are attempting to do is develop a vocabulary of just such individually recognizable part models. One may, therefore, think of the research described here as returning to the blocks world, but with models of 3-D structure that are tremendously more sophisticated than simple blocks or polyhedra.

We believe that the modeling language presented here has a good chance of being able to handle most of the forms found in the real world. The images in this paper demonstrate the expressive power of this new vocabulary of models (their cartoon-like nature is primarily due to the lack of surface texturing), and the mathematics in this section of the paper demonstrate the plausibility of recovering such part descriptions from sparse and partial image data. Even if it should turn out that our models aren't yet sophisticated enough to deal with the complexity of real world, we will have *at least* made major progress towards bridging the gap between the present state of the art and that needed to construct a general-purpose, real-world vision system.

## 5. Using the Representation

The particular models of the world that perception uses to interpret sensory data induce a profound organization on all of our conceptual structures. If we stand in the center of Stonehenge, we can see either a collection of pillars, several irregular walls of pillars, or concentric circular structures with regularly spaced pillars. This is the familiar Gestalt phenomenon of grouping; what is important about it is that *which* grouping you spontaneously see strongly influences what hypotheses you entertain when trying to deduce, for instance,

the purpose of Stonehenge. Examples such as this demonstrate that the manner in which perception "carves up" the world—that is, its models of the world—strongly determine the way in which we think about the world.

The issue of perceptual models is, therefore, of more than passing interest to those interested in cognition. It seems reasonable that if we are to develop machines that are able to display common-sense reasoning abilities, for instance, we must have spatial representations that are at least roughly equivalent to those people employ in organizing their picture of the world. Similarly, if we are ever to communicate with machines about our shared environment we must develop spatial representations that are at least isomorphic to the representations that we use. We must have a representation that captures the same sorts of distinctions we make when we carve objects into parts.

Because communication depends upon having a shared representation of the situation, we can use man–machine communication as a fairly sensitive test of whether a particular representation captures the notions of difference and similarity that humans employ. The empirical (and so far informal) finding that the organization of our shape descriptions correspond closely with the human perceptual organization is, as a consequence, quite interesting: the representation seems to offer exciting possibilities for flexible, effective man–machine communication. It was, therefore, of great interest to test how effectively we can use the representation described here as a basis for communication between a computer and its operator concerning image data and 3-D shape.

### 5.1. Communicating about a digital terrain map

As a first experiment we took the problem of communicating with a computer about a digital terrain map, as might be done in guiding a stereo compilation process or when plotting a path through the terrain. Fig. 7 showed how a mountain-like surface can be built up from the combination of progressively smaller primitives. We can also take a real surface, such as the digital terrain map of Yosemite Valley shown in Fig. 10(a) and decompose it into a canonical lump description by use of a minimum-complexity criterion, that is, we attempt to account for the shape with the fewest number of component parts as is possible (see [49]). One simple mechanism for approximating this decomposition is to form a Laplacian pyramid [50], examine the entries in this pyramid to find those points that most closely correspond to the shape of a single "lump" (by looking at the neighbors of the point in both space and scale), subtract off that lump from the original form. We then repeat this procedure until no entries remain in the pyramid.

If we want to have a "sketch" of the DTM surface (a simplified description that we can use for communication), we can use estimates of the surface's variance and fractal dimension to set an acceptance threshold, so that our

decomposition procedure finishes by taking only the 50 or so most prominent surface features. To adequately characterize a DTM we have found that we need to look for only two types of primitive elements: one, a vertically oriented symmetrical peak, and two, a horizontally oriented elongated ridge or valley. The fractal statistics of the surface characterize how features of the surface change across scale and, therefore, gives us the information needed to adjust the acceptance threshold for different scales, so that the prominence of features

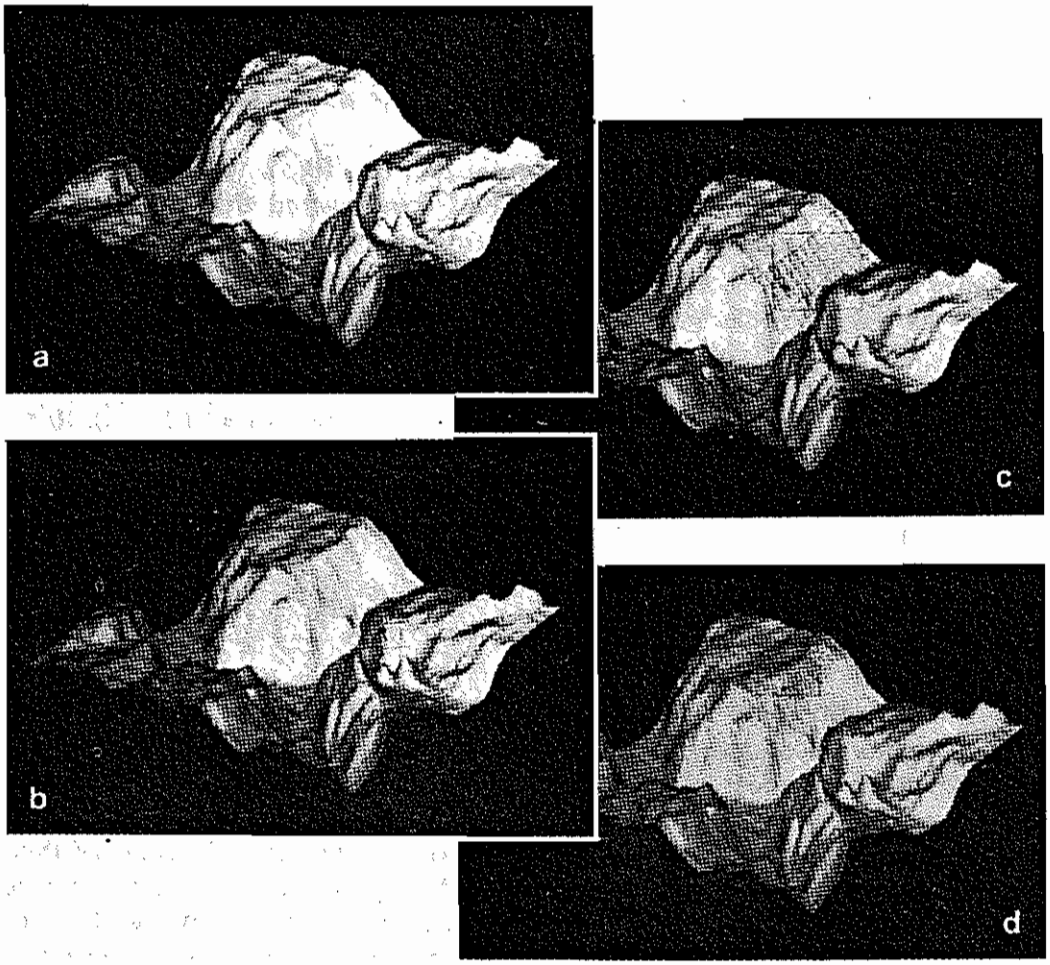



FIG. 10. (a) A digital terrain map of Yosemite Valley, which is automatically decomposed into a "sketch," a description in our representational system that contains terms ("lumps") that correspond roughly to "peaks," "valleys," and "ridges," so that the parts of this description correspond closely with the perceptual organization that we impose on the scene. This is illustrated in (b), (c) and (d), which show a person pointing to a part of the image, and the computer using this sketch to determine what part of the terrain is being gestured at, and highlighting the "part" referred to by covering it with cross-hatching. This decomposition of the scene into perceptually salient "parts" thus fulfills a critical requirement for effective man–machine communication: similar representations of the scene.

accepted at one scale corresponds to the prominence of smaller- or larger-scale features. When we do this, the result is a description that organizes the pixel data into its most prominent components at all scales, in a way that we have found corresponds closely with our naive perceptual organization of the surface—e.g., organizing the surface into peaks ridges, valleys, and the like.

The ability to structure the pixel data in a manner that corresponds to the perceptual organization we impose upon the data allows us to support human–computer communication about the scene. It allows us to point to a part of the scene, say "that thing" and have the machine be able to make a good guess about what part of the surface we want to indicate, as opposed to the current state of the art in which we have to carefully outline the part of the surface that we want manipulate.

This sort of communication is illustrated in Figs. 10(h), (c) and (d), which shows the operation of a program we have constructed that performs this parsing of a digital terrain map (DTM), identifies the 50 or so most prominent perceptual "parts," and then allows the user to interact with the DTM by simply pointing to peaks, valleys, ridges and so forth. These figures show a user pointing, the program interpreting what "feature" the user intended to indicate, and then highlighting that feature by cross-hatching it. The highlighted feature can then be edited to improve the DTM, defined as a primitive object in a path planning calculation, or used in whatever manner the user's purpose demands. As these figures illustrate, we have found a good correspondence between this program's structuring of the image and the structure people impose on the image.

### 5.2. Building 3-D models

One other example that illustrates using the representation to facilitate man–machine communication is the 3-D modeling system called "SuperSketch" that was used to make most of the images in this paper. In this Symbolics-3600-based modeling system users create "lumps," change their squareness/roundness, stretch, bend, and taper them, and make Boolean combinations of them in real time by moving the mouse through the relevant parameter space, controlling which parameter is being varied by using the mouse buttons. Because these forms have an underlying analytical form, we can use fast, qualitative approximations to accomplish hidden-surface removal, intersection and image-intensity calculations in real time—something that could not be accomplished on a Symbolics 3600 if a polygon-based description were employed. "Real time" in this case means that a "lump" can be moved, hidden-surface removal accomplished, and drawn as a 200-polygon line drawing-approximation in one eighth of a second, and a complex, full color image such as Fig. 1 can be rendered in approximately 20 seconds.

Because the primitives, operations and combining rules used by the com-

puter are very well matched to those of the human operator, we have found that interaction is surprisingly effortless: it took a relatively unskilled operator less than a half hour to assemble the face in Fig. 6, about ten minutes to create the lobster in Fig. 3, and about four hours total to make Fig. 1. This is in rather stark contrast to more traditional 3-D modeling systems that might require days or weeks to build up a scene such as shown in Fig. 1. This performance, perhaps more than any other statistic that could be given, illustrates how the close match between this representational system and the perceptual organization employed by human operators facilitates effective man–machine communication.

## 6. Summary

To support our reasoning abilities perception must recover environmental regularities—e.g., rigidity, "objectness," axes of symmetry—for later use in cognitive processes. Understanding this recovery of structure is critically important because the structural organization that perception delivers to cognition is the foundation upon which we construct our picture of the world; these regularities are the building blocks of all cognitive activities.

To create a theory of how our perceptual apparatus can produce meaningful cognitive building blocks from an array of image intensities we require a representation whose elements may be lawfully related to important physical regularities, and that correctly describes the perceptual organization people impose on the stimulus. Unfortunately, the representations that are currently available were originally developed for other purposes (e.g., physics, engineering) and have so far proven unsuitable for the problems of perception or common-sense reasoning.

For instance, the complexity of standard descriptions for such common natural forms as clouds, human faces, or trees has been a fundamental block to progress in computational psychology, artificial intelligence and machine vision. It is a fundamental result of mathematics that one cannot recover 3-D shape descriptions from an image when the number of parameters to be recovered is greater than the number of pixels in the image. How, then, can we hope to understand perception when our representational tools force us into the uncomfortable position of knowing a priori that we cannot recover the desired descriptions from image data? Further, even if we *could* recover such descriptions, how can we hope to understand common-sense reasoning if forced to use such overly complex descriptions?

In answer to these problems we have presented a representation that has proven competent to accurately describe an extensive variety of natural forms (e.g., people, mountains, clouds, trees), as well as man-made forms, in a succinct and natural manner. The approach taken in this representational system is to describe scene structure at a scale that is more like our naive

perceptual notion of "a part" than the pointwise descriptions typical of current image understanding research, and to use a description that reflects a possible formative history of the object, e.g., how the object might have been constructed from lumps of clay.

Each of the component parts of this representation—superquadric "lumps," deformations, Boolean combination, and the recursive fractal construction—have been previously suggested as elements of various shape descriptions, usually for other purposes. The contribution of this paper is to bring all of these separate descriptive elements together, and employ them as a representation for natural forms and as a theory of perceptual organization. In particular, we believe that the important contributions of this paper are the following.

(1) We have demonstrated that this process-oriented representational system is able to accurately describe a very wide range of natural and man-made forms in an extremely simple, and therefore useful, manner. Further, the representation can be used to support fast, qualitative approximations to determine, e.g., intersection, appearance or relative position. Such qualitative reasoning is employed in SuperSketch allow real-time movement, deformation, Boolean combination, hidden-surface removal, intersection and rendering.

(2) We have found that descriptions couched in this representation are similar to people's (naive) verbal descriptions and appear to match people's (naive) perceptual notion of "a part", this correspondence is strong evidence that the descriptions we form will be good spatial primitives for a theory of common-sense reasoning. Additionally, we hope that this descriptive system will provide the beginnings of a rigorous, mathematical treatment of the still vaguely defined subject of human perceptual organization.

(3) The part-model approach to perception makes the problem of recovering shape descriptions overconstrained and therefore potentially extremely reliable, while still providing the flexibility to learn new object descriptions. Toward this end we have shown that our current descriptive vocabulary is capable of describing a wide range of natural forms, and that the primitive elements of this language can be recovered from partial image data in an overconstrained and apparently noise-insensitive manner.

(4) And finally, we have shown that descriptions framed in the representation have markedly facilitated man–machine communication about both natural and man-made 3-D structures. It appears, therefore, that this representation gives us the right "control knobs" for discussing and manipulating 3-D forms.

The representational framework presented here is *not* complete. It seems clear that additional process-oriented modeling primitives, such as branching structures [21] or particle systems [51], will be required to accurately represent objects such as trees, hair, fire, or river rapids. Further, it seems clear that domain experts form descriptions differently than naive observers, reflecting their deeper understanding of the domain-specific formative processes and

their more specific, limited purposes. Thus, accounting for expert descriptions will require additional, more specialized models. Nonetheless, we believe this descriptive system makes an important contribution toward solving current problems in perceiving and reasoning about natural forms, by allowing us to construct accurate models that are still simple enough to be useful, and by providing us with the basis for more effective man–machine communication.

## ACKNOWLEDGMENT

I would like to thank Andy Witkin for his collaboration in developing the basic elements of this approach to perception, and for his help in refining my ideas about fractals and superquadrics. I would also like to thank David Heeger, Oscar Firschein, Bob Bolles and Tracy Heibeck for their help and criticism in the writing of this manuscript.

This research was made possible by National Science Foundation, Grant No. DCR-83-12766, by Defense Advanced Research Projects Ageney contract no. MDA 903-83-C-0027, and by a grant from the systems Development Foundation.

## REFERENCES

1. Thompson, D'Arcy, *On Growth and Form* (University Press, Cambridge, U.K., 2nd ed., 1942).
2. Stevens, S., *Patterns in Nature* (Atlantic-Little, Brown Books, Boston, MA, 1974).
3. Rosch, E., On the internal structure of perceptual and semantic categories, in: T.E. Moore (Ed.), *Cognitive Development and the Acquisition of Language* (Academic Press, New York, 1973).
4. Wertheimer, M., Laws of organization in perceptual forms, in: W.D. Ellis (Ed.), *A Source Book of Gestalt Psychology* (Harcourt Brace, New York, 1923).
5. Johansson, G., *Configurations in Event Perception* (Almqvist and Wiksell, Stockholm, 1950).
6. Marr, D. and Nishihara, K., Representation and recognition of the spatial organization of three-dimensional shapes, *Proc. Roy. Soc. London B* **200** (1978) 269-294.
7. Nishihara, H.K., Intensity, visible-surface and volumetric representations. *Artificial Intelligence* **17** (1981) 265–284.
8. Binford, T.O., Visual perception by computer, in: *Proceedings IEEE Conference on Systems and Control*, Miami, FL, 1971.
9. Gibson, J.J., *The Ecological Approach to Visual Perception* (Houghton Mifflin, Boston, MA, 1979).
10. Marr, D. *Vision* (Freeman, San Francisco, CA, 1982).
11. Agin, G.J. and Binford, T.O., Computer descriptions of curved objects, *IEEE Trans. Comput.* **25** (1976) 439–449.
12. Nevatia, R. and Binford, T.O., Description and recognition of curved objects *Artificial Intelligence* **8** (1977) 77–98.
13. Badler, N. and Bajacsy, R., Three-dimensional representations for computer graphics and computer vision, *Comput. Graphics* **12** (1978) 153–160.
14. Brady, J.M., Describing visible surfaces, in: A. Hanson and E. Riseman (Eds.), *Computer Vision Systems* (Academic Press, New York, 1978).
15. Brooks, R., Model based 3-D interpretation of 2-D images, in: A. Pentland (Ed.), *From Pixels to Predicates* (Ablex, Norwood, NJ, 1985).
16. Bolles, B. and Haroud, R., 3DPO: An inspection system, in A. Pentland (Ed.), *From Pixels to Predicates* (Ablex, Norwood, NJ, 1985).
17. Barrow, H.G. and Tenebaum, J.M., Recovering intrinsic scene characteristics from images, in: A. Hanson and E. Riseman (Eds.), *Computer Vision Systems* (Academic Press, New York, 1978).

18. Pentland, A., Local analysis of the image, *IEEE Trans. Pattern Anal. Machine Intelligence* **6** (1984) 170–187.
19. Witkin, A.P. and Tenebaum, J.M., On perceptual organization, in: A. Pentland (Ed.), *From Pixels to Predicates* (Ablex, Norwood, NJ, 1985).
20. A. Pentland and A. Witkin, On perceptual organization, in: *Proceedings Second Conference on* Perceptual Organization, Pajaro Dunes, CA, 1984.
21. Smith, A.R., Plants, fractals and formal languages, *Comput. Graphics* **18** (3) (1984) 1–11.
22. Mandelbrot, B.B., *The Fractal Geometry of Nature* (Freeman, San Francisco, CA, 1982).
23. Georgeff, M.P. and Wallace, C.S., A general selection criterion for inductive inference, in: *Proceedings Sixth European Conference on Artificial Intelligence*, Pisa, Italy, 1984.
24. Herman, M. and Kanade, T., The 3-D mosaic scene understanding systems, in: A. Pentland (Ed.), *From Pixels to Predicates* (Ablex, Norwood, NJ, 1985).
25. Konderink, J.J. and Van Doorn, A.J., The shape of smooth objects and the way contours end, *Perception* **11** (1982) 129–137.
26. Konderink, J.J. and Van Doorn, A.J., The internal representation of solid shape with respect to vision, *Biol. Cybernet.* **32** (1979) 211–216.
27. Hoffman, D. and Richards, W., Parts of recognition, in: A. Pentland (Ed.), *From Pixels to Predicates* (Ablex, Norwood, NJ, 1985).
28. Barr, A., Superquadrics and angle-preserving transformations, *IEEE Comput. Graphics Appl.* **1** (1981) 1–20.
29. Kauth, R., Pentland, A. and Thomas, G., BLOB: an unsupervised clustering approach to spatial grouping in: *Proceedings Eleventh International Symposium on Remote Sensing of the Environment*, Ann Arbor, MI 1977.
30. Hobbs, J. Final Report on Commonsense Summer, Tech. Note 370, SRI Artificial Intelligence Center, Menlo Park, CA, 1985.
31. Barr, A., Global and local deformations of solid primitives, *Comput. Graphics* **18** (3) (1984) 21–30.
32. Hollerbach, J.M., Hierarchical shape description of objects by selection and modification of prototypes, Ph.D. Theses, AI Tech. Rept. 346, MIT, Cambridge, MA, 1975.
33. Hayes, P., The second naive physics manifesto, in: J. Hobbs and R. Moore (Eds.), *Formal Theories of the Commonsense World* (Ablex, Norwood, NJ, 1985).
34. Pentland, A., Fractal-based description of natural scenes, *IEEE Trans. Pattern Anal. Machine Intelligence* **6** (1984) 661–674.
35. Pentland, A., Fractal-based description in: *Proceedings Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, F.R.G. (1983) 973–981.
36. Medioni, G. and Yasumoto, Y., A note on using the fractal dimension for segmentation, *IEEE Computer Vision Workshop*, Annapolis, MD, 1984.
37. Pentland, A., Shading into texture in: *Proceedings Fourth National Conference on Artificial Intelligence*, Austin, TX (1984) 269–273.
38. Pentland, A., Fractals: a model for both texture and shading, *Optic News* (October, 1984) 71.
39. Pentland, A., Perception of three-dimensional textures, *Investigative Opthomology and Visual Science* **25** (3) (1984) 201.
40. Fodor, J., *Modularity of Mind: An Essay on Faculty Psychology* (MIT Press, Cambridge MA, 1982).
41. Gregory, R.L., *The Intelligent Eye* (McGraw-Hill, New York, 1970).
42. Leyton, M., Perceptual organization as nested control, *Biol. Cybernet.* **51** (1984) 141–153.
43. Held, R. and Richards, W., (Eds.) *Recent Progress in Perception, Readings from Scientific American* (Freeman, San Francisco, CA, 1975).
44. Ferrie, F.P. and Levine, M.D., Piecing together the 3-D shape of moving objects: an overview, in: *Proceedings IEEE Conference on Vision and Pattern Recognition*, San Francisco, CA, 1985.
45. Roberts, L., Machine perception of three-dimensional solids, in: J.T. Tippet, et al. (Eds.), *Optical and Electrooptical Information Processing* (MIT Press, Cambridge, MA, 1965).

46. Winston, P.H., Learning structural descriptions from examples, in: P.H. Winston (Ed.), *The Psychology of Computer Vision* (McGraw-Hill, New York, 1975).
47. Winston, P., Binford, T., Katz, B. and Lowry, M., Learning physical descriptions from functional definitions, examples, and precedents, in: *Proceedings Third National Conference on Artificial Intelligence*, Washington, DC (1983) 433–439.
48. Davis, E., The MERCATOR representation of spatial knowledge, in: *Proceedings Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, F.R.G. (1983) 295–301.
49. Pentland, A., On describing complex surfaces, *Image and Vision Computing* 3(4) (1985) 1–15.
50. Burt, P.J. and Adelson, E.H., The Laplacian pyramid as a compact image code, *IEEE Trans. Communications* **31** (1983) 532–540.
51. Reeves, W.T., Particle systems—a technique for modeling a class of fuzzy objects, *ACM Trans. Graphics* **2** (2) (1983) 91–108.

*Received August 1985*